Exploiting vulnerabilities of deep neural networks for privacy protection

Ricardo Sanchez-Matilla, Chau Yi Li, Ali Shahin Shamsabadi, Riccardo Mazzon, Andrea Cavallaro

***Abstract*—Adversarial perturbations can be added to images to protect their content from unwanted inferences. These perturbations may, however, be ineffective against classifiers that were not seen during the generation of the perturbation, or against defenses based on re-quantization, median filtering or JPEG compression. To address these limitations, we present an adversarial attack that is specifically designed to protect visual content against unseen classifiers and known defenses. We craft perturbations using an iterative process that is based on the Fast Gradient Signed Method and that randomly selects a classifier and a defense, at each iteration. This randomization prevents an undesirable overfitting to a specific classifier or defense. We validate the proposed attack in both targeted and untargeted settings on the private classes of the Places365-Standard dataset. Using ResNet18, ResNet50, AlexNet and DenseNet161 as classifiers, the performance of the proposed attack exceeds that of eleven state-of-the-art attacks.**

***Index Terms*—Deep learning, adversarial images, privacy protection.**

I. Introduction

Images shared online capture people and scenes that reveal personal information, as well as information about personal choices and preferences. This information can be automatically inferred by classifiers. To prevent this potential privacy violation and to protect the visual content from unwanted automatic inferences [1], we aim to exploit the vulnerability of classifiers to adversarial attacks [2], [3], [4].

An adversarial attack should mislead classifiers that the attacker has access to (*seen* classifiers), as well as classifiers that the attacker has no information about, not even the prediction output (*unseen* classifiers). However, adversarial attacks often fail to mislead unseen classifiers as the generated adversarial perturbations overfit to a specific classifier [5] or to an ensemble of classifiers [6]. Adversarial images should also be robust to *defenses*. Defenses can be based on re-quantization [7], median filtering [7] and JPEG compression [8]. Moreover, adversarial perturbations should not degrade the image quality, especially when added for protecting privacy [1], [9].

In this paper, we propose an adversarial attack that aims to prevent both seen and unseen classifiers from inferring private information present in an image, even when the classifiers are equipped with known defenses. At the core of the proposed attack there is an iterative combination of random selections of a classifier and a defense within a Fast Gradient Signed Method (FGSM) framework[1] [12]. This random selection avoids the creation of perturbations that overfit to a specific classifier or defense, and improves the misleading rate for seen and unseen classifiers. The proposed attack, which can work in both targeted and untargeted settings, is related to methods that are based on ensemble of classifiers [6] and on defense transformations [5], [13], [14], but differs in the fact that both classifiers and transformations are randomly chosen at each iteration. Moreover, the proposed attack enables the use of complex defense transformations with null derivative. We validate the proposed attack for the protection of scene content on the privacy subset of the Places365-Standard dataset [9], which consists of scenes such as places of worship and hospitals. We evaluate the attack on state-of-the-art classifiers, namely ResNet18, ResNet50, AlexNet, and DenseNet161, based on the misleading rate, detectability and image quality.

II. Problem definition

Let $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$ be an image of width W and height H pixels, and C color channels whose dynamic ranges are $[0, 255]$. Let $M(\cdot)$ be a D-class deep neural network classifier with parameters θ and trained using the cost function $J_M(\cdot)$.

Let $\hat{y}_{\mathbf{x}}$ be the *true* class associated with $\mathbf{x}$. The classifier outputs a prediction vector for $\mathbf{x}$, $\mathbf{p}_{\mathbf{x}} = (p_i)_{i=1}^{D} = M(\mathbf{x})$, where p_i represents the probability of $\mathbf{x}$ being associated with class $i \in \{1, ..., D\}$. The *predicted* class for $\mathbf{x}$, $y_{\mathbf{x}}$, is the most likely of D classes:

$$y_{\mathbf{x}} = \underset{i=1,...,D}{\arg\max}\ p_i. \tag{1}$$

Note that the predicted class might differ from the true class, which is unknown during the execution of the adversarial attack. Adversarial attacks for privacy protection should aim to hide the true class, $\hat{y}_{\mathbf{x}}$, from $M(\cdot)$, even when the predicted class is incorrect. The adversarial perturbation, $\boldsymbol{\delta}_{\mathbf{x}} \in \mathbb{R}^{W \times H \times C}$, added to the original image, $\mathbf{x}$, generates an adversarial image, as $\dot{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}_{\mathbf{x}}$. This perturbation causes the classifier to predict an *adversarial* class, $y_{\dot{\mathbf{x}}}$, by decreasing the probability of the predicted class, $y_{\mathbf{x}}$, (untargeted attack [12], [13], [15], [16]) until

$$y_{\dot{\mathbf{x}}} \neq y_{\mathbf{x}}, \tag{2}$$

[1]FGSM is the iterative attack proposed by Kurakin *et al.* [10]. Note that this attack is also known as Basic Iterative Method (BIM) [10] and as Projected Gradient Descent (PGD) attack with L_∞ norm [11].

Manuscript received May 5, 2019; revised December 14, 2019 and March 24, 2020; accepted March 26, 2020. *(Chau Yi Li and Ali Shahin Shamsabadi equally contributed).* The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Wang. *(Corresponding author: Ricardo Sanchez-Matilla).*

The authors are with the Centre for Intelligent Sensing (CIS), Queen Mary University of London (QMUL), London E1 4NS, U.K. (e-mail: ricardo.sanchezmatilla@qmul.ac.uk; chauyi.li@qmul.ac.uk; a.shahinshamsabadi@qmul.ac.uk; r.mazzon@qmul.ac.uk; a.cavallaro@qmul.ac.uk

Digital Object Identifier 10.1109/TMM.2020.2987694

TABLE I
COMPARISON OF ADVERSARIAL ATTACKS: CHOICE OF NORM AND CORRESPONDING DENSITY OF THE PERTURBATION, WHICH CAN BE SPARSE (S) OR DENSE (D). THE CLASS SELECTION (TYPE) CAN BE UNTARGETED (U) OR TARGETED (T). AN ATTACK CAN BE DESIGNED FOR MULTIPLE CLASSIFIERS (MC) AND TO WITHSTAND IMAGE TRANSFORMATIONS (TRANS.).

Reference	Attack	Norm	Perturbation	Type	MC	Trans.
[3]	JSMA	L_0	S	T		
[4]	CW	L_0, L_2, L_∞	S, D	T		
[15]	DeepFool	L_2	D	U		
[16]	SparseFool	L_1	S	U		
[12]	U-FGSM	L_∞	D	U		
[10]	R-FGSM	L_∞	D	T		
[12]	L-FGSM	L_∞	D	T		
[1]	P-FGSM	L_∞	D	T		
[5]	EOT	L_∞	D	T		✓
[13]	DI-FGSM	L_∞	D	U, T	✓	✓
[6]	E-FGSM	L_∞	D	T	✓	
Proposed	RP-FGSM	L_∞	D	U, T	✓	✓

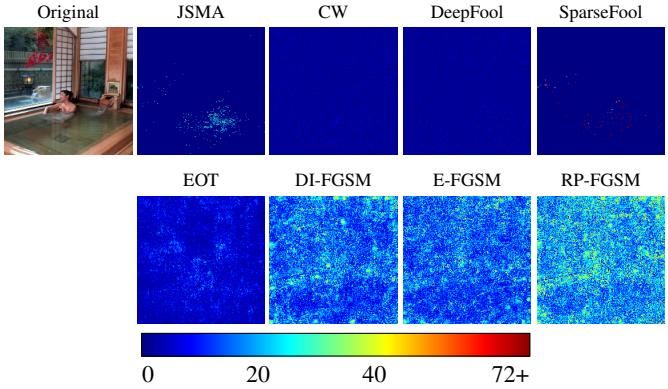


Fig. 1. Cumulative adversarial perturbations (sum of the absolute difference between original and adversarial images across the three color channels) generated by JSMA, CW, DeepFool, SparseFool, EOT, DI-FGSM (untargeted), E-FGSM, and the proposed RP-FGSM (untargeted).

or by increasing the probability of a specific target class, y_t, (targeted attack [1], [3], [4], [6], [10], [12], [13]), such that

$$y_{\dot{\mathbf{x}}} = y_t \neq y_{\mathbf{x}}. \tag{3}$$

The target class can be determined randomly [1], [12], systematically as the least-likely class [12], or adaptively from the prediction vector [1].

Defenses against adversarial attacks aim to eliminate, prior to inputting images to the classifier, the effect of possible adversarial perturbations using a transformation, $\phi(\cdot)$, namely median filtering [7], re-quantization [7] or JPEG compression [8]. Moreover, to detect an image as adversarial, the probability vector from a classifier for an image $\dot{\mathbf{x}}$ and its transformed version, $\phi(\dot{\mathbf{x}})$, can be compared with the L_1 norm as

$$\|M(\dot{\mathbf{x}}) - M(\phi(\dot{\mathbf{x}}))\|_1 > \tau, \tag{4}$$

where τ is learned to accept a specific false-positive rate [7].

III. BACKGROUND

Adversarial images are crafted by constraining the added perturbation with, typically, an L_p norm between $\mathbf{x}$ and $\dot{\mathbf{x}}$. Attacks may constrain the total number of perturbed pixels (L_0) [3], the sum of magnitudes (L_1) [16], the Euclidean distance (L_2) [4], [15], or the maximum per-pixel variation (L_∞) [1], [6], [12]. As the optimization of the perturbation, regularized by the norm, has no closed-form solution due to non-linear operations with the parameters, θ, and non-convex cost functions used to train the deep neural network classifiers, several adversarial attacks iteratively generate the adversarial image, $\dot{\mathbf{x}}$, as [12]:

$$\dot{\mathbf{x}}_{n+1} = \dot{\mathbf{x}}_n + \boldsymbol{\delta}_{\dot{\mathbf{x}}_n}, \tag{5}$$

from $\dot{\mathbf{x}}_0 = \mathbf{x}$. The iteration process stops when a specific number of iterations, N, is reached, i.e. $\dot{\mathbf{x}} = \dot{\mathbf{x}}_{N+1}$, or when the misleading objective is achieved [3], [15]. N is typically chosen as a function of parameters linked to the preservation of image quality [3], [10]. Table I provides a comparative summary of adversarial attacks. Moreover, Figure 1 shows examples of the magnitude of perturbations generated by representative adversarial attacks, which are described next.

JSMA [3], a targeted attack, uses the L_0 norm and aims to identify and perturb by one intensity unit, at each iteration, the *two* most effective pixels. JSMA increases the probability of the target class, y_t, compared with that of any other classes, using a saliency score, $S(\cdot)$, which determines the effect of perturbing a pixel, $x_i \in \mathbf{x}$:

$$S(x_i, y_t) = \tag{6}$$
$$\begin{cases} 0 & \text{if } \nabla_{x_i} M_t(\mathbf{x}) < 0 \text{ or } \sum_{j \neq t} \nabla_{x_i} M_j(\mathbf{x}) > 0 \\ \nabla_{x_i} M_t(\mathbf{x}) \left\| \sum_{j \neq t} \nabla_{x_i} M_j(\mathbf{x}) \right\|_1 & \text{otherwise,} \end{cases}$$

where $\nabla_{x_i} M_t(\mathbf{x})$ is the gradient of the target class, y_t, with respect to pixel x_i. The attack iteratively generates the perturbation, until the adversarial image is classified as the target class or a specified number of pixels have been perturbed.

Carlini-Wagner (CW) [4], a targeted attack, maximizes the difference between the logarithmic probabilities of the target class and other classes using the L_2 norm. Three forms of selecting the target classes were used, namely uniform random selection, second- and least-likely cases.

DeepFool [15], an untargeted attack, crafts adversarial perturbations controlled by the L_2 norm. At each iteration, the adversarial perturbation is the orthogonal projection of the adversarial image from the previous iteration onto the closest linearized class boundary of $M(\cdot)$. The final adversarial image, $\dot{\mathbf{x}} = \dot{\mathbf{x}}_{N+1}$, is the one that exceeds the closest class boundary as

$$\dot{\mathbf{x}}_{N+1} = \dot{\mathbf{x}}_N + (1+\eta)\boldsymbol{\delta}_{\dot{\mathbf{x}}_N}, \tag{7}$$

where $\eta \ll 1$ and $\boldsymbol{\delta}_{\dot{\mathbf{x}}_N}$ is the final adversarial perturbation. As DeepFool might generate adversarial images whose pixels exceed the allowed dynamic range, SparseFool [16], an untargeted attack similar to DeepFool, clips the pixel values within $[\mathbf{l}, \mathbf{u}]$, as

$$\min \|\dot{\mathbf{x}} - \mathbf{x}\|_1 \quad \text{s.t.} \quad y_{\dot{\mathbf{x}}} \neq y_{\mathbf{x}} \quad \text{and} \quad \mathbf{l} \leq \dot{\mathbf{x}} \leq \mathbf{u}. \tag{8}$$

This minimization is solved based on the low-mean curvature properties of the decision boundary of each image [17]. Note

that, unlike DeepFool, SparseFool uses the L_1 norm, thus generating sparse perturbations.

The family of attacks based on FGSM generates the perturbation in the direction of the cost function, $J_M(\cdot)$, in order to maximize the cost of remaining in the predicted class $y_{\mathbf{x}}$ (untargeted) or to minimize the cost of predicting the targeted class y_t (targeted):

$$\boldsymbol{\delta}_{\mathbf{x}} = \begin{cases} \delta \operatorname{sign}\left(\nabla_{\mathbf{x}} J_M\left(\theta, \mathbf{x}, y_{\mathbf{x}}\right)\right) & \text{untargeted attack,} \\ -\delta \operatorname{sign}\left(\nabla_{\mathbf{x}} J_M\left(\theta, \mathbf{x}, y_t\right)\right) & \text{targeted attack,} \end{cases} \tag{9}$$

where δ controls the magnitude of the perturbation, $\operatorname{sign}(\cdot)$ is the sign function and $\nabla_{\mathbf{x}} J_M(\cdot)$ is the gradient of the cost function, $J_M(\cdot)$, showing the updated direction with respect to $\mathbf{x}$. FGSM-based attacks use the L_∞ norm and limit $\dot{\mathbf{x}}$ to the ϵ-neighborhood of $\mathbf{x}$ and within the dynamic range of an image at each iteration, as

$$\dot{\mathbf{x}}_{n+1} = \mathcal{C}_{\mathbf{x},\epsilon}\left(\dot{\mathbf{x}}_n + \boldsymbol{\delta}_{\dot{\mathbf{x}}_n}\right), \tag{10}$$

where $\mathcal{C}_{\mathbf{x},\epsilon}(\cdot)$ is a clipping function, defined as

$$\mathcal{C}_{\mathbf{x},\epsilon}(\dot{\mathbf{x}}) = \min\big\{255, \mathbf{x} + E, \max\left\{0, \mathbf{x} - E, \dot{\mathbf{x}}\right\}\big\}, \tag{11}$$

where $E = \{\epsilon\}^{W \times H \times C}$. The value of ϵ is a trade-off between the misleading rate and the quality of the adversarial image: the larger ϵ, the higher the potential misleading rate but, also, the stronger the image degradation.

To evade defenses based on transformations and to preserve visual quality, Expectation Over Transformation (EOT) [5] optimizes the loss on the target class over a set of pre-defined 2D transformations, $\boldsymbol{\Phi}_{2D}$, while minimizing the distance between the transformed original image and the transformed adversarial image in the Lab color space [18]:

$$\begin{aligned} \dot{\mathbf{x}}_{n+1} = & \mathcal{C}_{\mathbf{x},\epsilon}(\dot{\mathbf{x}}_n - \delta \operatorname{sign}(\nabla_{\mathbf{x}} J_M(\theta, \phi_{2D}(\dot{\mathbf{x}}_n), y_t) + \\ & + \lambda \|\mathcal{L}(\phi_{2D}(\dot{\mathbf{x}}_n)) - \mathcal{L}(\phi_{2D}(\dot{\mathbf{x}}_n))\|_2), \end{aligned} \tag{12}$$

where $\mathcal{L}(\cdot)$ is a function that converts images from the RGB to the Lab color space, λ controls the visual similarity, and $\phi_{2D}(\cdot) \in \boldsymbol{\Phi}_{2D}$ is an image transformation chosen with uniform random probability at each iteration.

To increase the misleading rate with unseen classifiers, similarly to EOT, Diverse Input Fast Gradient Sign Method (DI-FGSM) [13] applies a random resizing followed by a padding transformation, with a pre-defined probability, on the adversarial image at each iteration.

To improve misleading performance with unseen classifiers, Ensemble FGSM (E-FGSM) [6] employs multiple classifiers, $M_k(\cdot)$, simultaneously when creating the perturbation:

$$\dot{\mathbf{x}}_{n+1} = \mathcal{C}_{\mathbf{x},\epsilon}\left(\dot{\mathbf{x}}_n - \delta \operatorname{sign}\left(\sum_{k=1}^{K} \nabla_{\mathbf{x}} J_{M_k}\left(\theta, \dot{\mathbf{x}}_n, y_t\right)\right)\right), \tag{13}$$

where $K \geq 1$. However, the use of all available classifiers at each iteration results in an overfitting that limits the ability of the adversarial image to mislead unseen classifiers [13].

Regarding the approach for selecting the target class, Least-likely FGSM (L-FGSM) [12] selects the least-likely predicted class. However, this systematic target class selection can compromise the protection of the image, as the selection process can be reversed [1]. Random-FGSM (R-FGSM) [10], E-FGSM [6] and EOT [5] randomly select the target class from all possible classes except the predicted class and the risk of reversibility is negligible. However, as the goal of the adversarial attacks is to hide the true class, which may not be the same as the predicted class (see Table II), *this strategy can result in the selection of the true class as the target*. Private-FGSM (P-FGSM) [1] avoids targeting the true class by discarding from the selection process the top predicted classes, which are more likely to contain the true class. The selection from the remaining classes is still random, thus limiting the risk of reversibility. Instead, untargeted approaches [12], [13] do not need the selection of a target class. However, knowledge of the true class is required for privacy protection.

TABLE II
CLASSIFICATION ACCURACY IN THE TEST SET OF THE PRIVATE PLACES365 DATASET FOR RESNET18, RESNET50, ALEXNET AND DENSENET161. KEY - T1: TOP-1 CLASSIFICATION ACCURACY; T5: TOP-5 CLASSIFICATION ACCURACY.

ResNet18		ResNet50		AlexNet		DenseNet161	
T1	T5	T1	T5	T1	T5	T1	T5
54.6	84.4	56.4	86.5	47.7	79.0	58.4	86.6

IV. ROBUST AND PRIVATE FGSM

We aim to generate adversarial perturbations that protect the *true* class of images and thus protect the private information they contain from unwanted automatic inferences. These adversarial perturbations should be robust to defenses, be undetectable, preserve image quality, and mislead both seen and unseen classifiers.

We propose an iterative approach, *Robust Private FGSM* (RP-FGSM), that avoids overfitting by randomly selecting, at each iteration, a classifier to attack and a defense to evade. The proposed approach differs from E-FGSM [6] and DI-FGSM [13], which consider an ensemble of classifiers, and from EOT [5] and DI-FGSM [13], which use only non-defense transformations (e.g. rotations). Moreover, as in other FGSM-based attacks [12], RP-FGSM maintains image quality inherently by controlling the magnitude of the perturbation with the parameter ϵ. The block diagram of RP-FGSM is shown in Figure 2.

Let $\mathbf{M} = \{M_k(\cdot)\}_{k=1}^{K}$ be a set of $K \geq 1$ classifiers, and $\boldsymbol{\Phi} = \{\phi_f(\cdot)\}_{f=0}^{F}$ be a set of F defense *transformations*, where $\phi_0(\cdot)$ is the identity function (i.e. no transformation is applied to the input). For a classifier $M_k(\cdot)$ and transformation $\phi_f(\cdot)$, let $y_{\dot{\mathbf{x}}}^k$ and $y_{\phi_f(\dot{\mathbf{x}})}^k$ be the predicted classes (Eq. 1) of the adversarial image, $\dot{\mathbf{x}}$, and the transformed adversarial image, $\phi_f(\dot{\mathbf{x}})$, respectively. We generate the adversarial image, $\dot{\mathbf{x}}$, for $\mathbf{x}$, whose true class is $\hat{y}_{\mathbf{x}}$, as

$$\hat{y}_{\mathbf{x}} \neq y_{\dot{\mathbf{x}}}^k \quad \text{and} \quad \hat{y}_{\mathbf{x}} \neq y_{\phi_f(\dot{\mathbf{x}})}^k \quad \forall f, k. \tag{14}$$

The process is initialized with $\dot{\mathbf{x}}_0 = \mathbf{x}$ and ends with $\dot{\mathbf{x}} = \dot{\mathbf{x}}_{N+1}$. At each iteration, n, we randomly select a transformation $R(\boldsymbol{\Phi}) \in \boldsymbol{\Phi}$ and a classifier $R(\mathbf{M}) \in \mathbf{M}$, where $R(\cdot)$ is a function that randomly selects an element from a set.

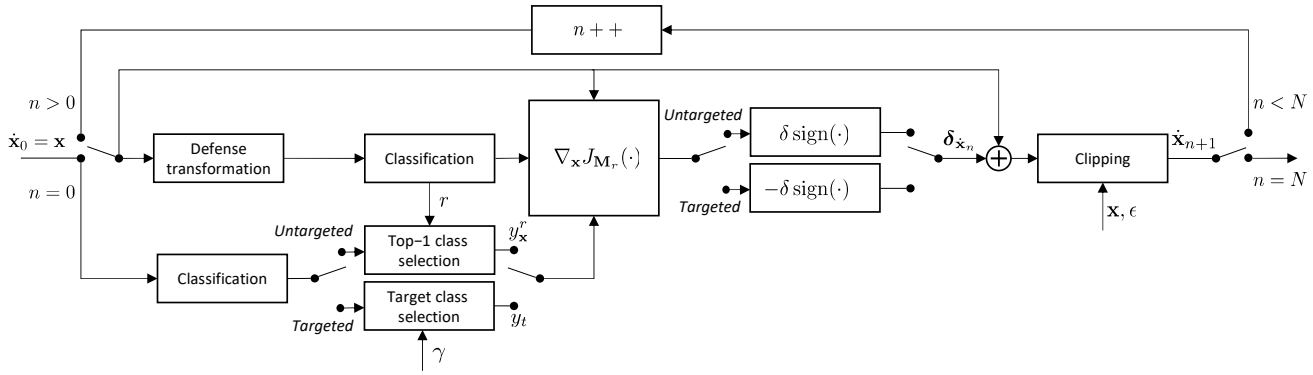


Fig. 2. Block diagram for the proposed adversarial attack, RP-FGSM. KEY – $\mathbf{x}$: original image; r: random index for selection of transformation and classifier; $\phi_r(\cdot)$: randomly chosen defense transformation; $M_r(\cdot)$: randomly chosen classifier; $\nabla_{\mathbf{x}} J_{M_r}$: gradient of the cost function $J(\cdot)$ for the classifier $M_r(\cdot)$ when predicting class y_t (targeted) or $y^r_{\mathbf{x}}$ (untargeted) with respect to $\mathbf{x}$; δ: magnitude of the perturbation added at each iteration; ϵ: parameter that controls the clipping function to maintain intensity values of each pixel within the ϵ-neighborhood of the original intensity value; γ: parameter for target class selection; N: number of iterations; $\dot{\mathbf{x}}_{n+1}$: adversarial image at iteration $n+1$. Note that when multiple classifiers are employed, the target class selection uses the intersection of the prediction vectors, as indicated in Eq. 17.

We use the most effective defense transformations, namely re-quantization [7], median filtering [7] and JPEG compression [8]. Both re-quantization and JPEG compression have a quantization and rounding step, where the derivative, needed to compute the perturbation, is zero for almost every pixel, thus resulting in a null adversarial perturbation. We prevent the generation of a null perturbation when applying these transformations by approximating each pixel of the image $\dot{\mathbf{x}}_n$ as $\lfloor \dot{\mathbf{x}}^i_n \rceil + (\dot{\mathbf{x}}^i_n - \lfloor \dot{\mathbf{x}}^i_n \rceil)^3$, where $\lfloor \cdot \rceil$ represents the rounding operator to the nearest integer, and i is the pixel index, whose derivative is unlikely to be zero [19]. To avoid overfitting to certain transformations and parameters, the parameters of each transformation (i.e. the number of bits for re-quantization, the kernel size for median filter and the compression parameters for JPEG compression) are also chosen randomly from a predefined set of values [13] (more details in Sec. V-B).

RP-FGSM can craft an adversarial image using targeted or untargeted attack by adding, at each iteration, the perturbation

$$\boldsymbol{\delta}_{\dot{\mathbf{x}}_n} = \begin{cases} \delta\ \mathrm{sign}(\nabla_{\mathbf{x}} J_{M_r}(\theta, \phi_r(\dot{\mathbf{x}}_n), y^r_{\mathbf{x}})) & \text{untargeted,} \\ -\delta\ \mathrm{sign}(\nabla_{\mathbf{x}} J_{M_r}(\theta, \phi_r(\dot{\mathbf{x}}_n), y_t)) & \text{targeted,} \end{cases} \tag{15}$$

where $y^r_{\mathbf{x}}$ is the class predicted by the classifier $M_r(\cdot)$. Moreover, RP-FGSM maintains the perturbation within the ϵ-neighborhood of $\dot{\mathbf{x}}$ as

$$\dot{\mathbf{x}}_{n+1} = \mathcal{C}_{\mathbf{x},\epsilon}\left(\dot{\mathbf{x}}_n + \boldsymbol{\delta}_{\dot{\mathbf{x}}_n}\right), \tag{16}$$

where $\mathcal{C}_{\mathbf{x},\epsilon}(\cdot)$ is the clipping function (Eq. 11).

Untargeted RP-FGSM crafts the adversarial perturbation in such a way that the predicted adversarial class moves farther away, in the decision space, from the most-likely predicted class for each classifier, which we assume to be in the neighborhood of the true class.

Targeted RP-FGSM selects the target class from a set that is more likely to exclude the true class, by leveraging the fact that the true class is often among the top classes predicted by a classifier [1]. Moreover, we reduce the risk that an attack is reversed by randomly selecting the target class. Specifically, we exploit the prediction probability vectors $\{\mathbf{p}^k_{\mathbf{x}}\}$ of the K classifiers. Let $\mathbf{p}'^k_{\mathbf{x}} = (p'^k_i)_{i=1}^D$ contain the elements of $\mathbf{p}^k_{\mathbf{x}}$ sorted in descending order. The target class is selected as

$$y_t = R\left(\bigcap_{k=1}^{K}\left\{y_{j+1} : \sum_{i=1}^{j} p'^k_i > \gamma, j \in \{1, ..., D-1\}\right\}\right), \tag{17}$$

where $R(\cdot)$ randomly selects an element from the set of candidate target classes, obtained as the intersection of the subset of classes whose cumulative probability exceeds a threshold $\gamma \in [0, 1]$. The larger γ, the fewer the classes. Figure 3 shows the influence of γ on the number of classes from which the target class y_t is selected.

After N iterations, the attack uses, on average, each classifier N/K times and each transformation N/F times. The number of iterations N might affect the quality of the image and therefore we choose it as a function of ϵ [10] (see Sec. V-B).

V. Validation

A. Experimental setup

We compare the targeted and untargeted versions of the proposed attack, Robust Private FGSM (RP-FGSM), with eleven state-of-the-art methods, namely Jacobian-based Saliency Map Approach (JSMA) [3], Carlini-Wagner (CW) with L_2 norm [4], DeepFool [15], SparseFool [16], iterative untargeted FGSM (U-FGSM) [12], Random FGSM (R-FGSM) [10], Least-likely FGSM (L-FGSM) [12], Private FGSM (P-FGSM) [1], Expectation Over Transformation (EOT) [5], Diverse Input FGSM (DI-FGSM, targeted and untargeted versions) [13] and Ensemble FGSM (E-FGSM) [6]. JSMA, CW, DeepFool and SparseFool are run with the parameters proposed by their authors. The target class for JSMA is the one that introduces the smallest perturbation, whereas for CW it is the second most-likely predicted class. The target class is the least-likely predicted class for L-FGSM, chosen as described in Eq. 17 for P-FGSM and RP-FGSM, and, for the other targeted FGSM-based attacks, is selected randomly from the set of possible classes, excluding the predicted class of the original image. For untargeted FGSM-based attacks, we use the predicted class to craft the adversarial perturbation.

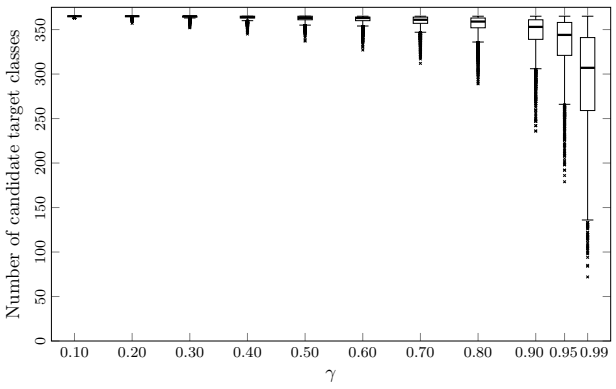


Fig. 3. Influence of γ on the number of candidate target classes, obtained as intersection of the subsets of classes whose cumulative probability exceeds the threshold γ. The plot is generated for 3000 images from the scene privacy subset defined in [9] using ResNet18, ResNet50 and AlexNet. The horizontal line within the box is the median; the lower and upper edges are the 25-percentile and 75-percentile, respectively. Each dot represents an outlier beyond 2.7 standard deviations from the average. When $\gamma = 0.1$, the median of the number of candidate classes is 364 (i.e. the target class is chosen from all but the predicted class for most images). As γ increases, the number of target-class candidates decreases. When $\gamma = 0.99$, the median decreases to 307 with a distribution that is more likely to exclude the true class from the target-class selection.

The Places365-Standard dataset [20], which has over 1.8 million images of 365 scene classes, is used for training the classifiers (see Table II). We use, as test set, the *scene privacy* subset of the validation set defined for the MediaEval 2018 Pixel Privacy Challenge [9], which comprises 60 private classes[2], that are a subset of the classes of the Places365-Standard dataset [20]. The test set is composed of 3,000 images, with 50 images from each of the (60) private classes. We consider four state-of-the-art classifiers, namely ResNet [21] with 18 and 50 layers, AlexNet [22] and DenseNet161 [23], trained for scene classification [20][3]. Table II reports their classification accuracy on the test set. Adversarial attacks are performed on ResNet18, ResNet50 and AlexNet, whereas DenseNet161, the most accurate classifier, is only used as an unseen classifier for testing. This represents the most challenging setup, given that the largest and most accurate classifier was never seen by the adversarial attacks [6].

[2]The subset of private classes includes scenes that may require, for various reasons, privacy protection, such as *army-base, bathroom, bedchamb, bedroom, church (indoor and outdoor), hospital and hospital-room, nursing-home, pharmacy, sauna, shower, swimming pool (indoor and outdoor), jacuzzi (indoor), temple (Asia)*; as well as scenes that would disclose private personal information, such as *airplane-cabin, airport terminal, amusement-park, aqueduct, bank-vault, bar, beach, beach-house, beer-garden, beer-hall, berth, bullring, bus-interior, bus station/indoor, campsite, car-interior, castle, catacomb, chalet, child-room, classroom, closet, coast, discotheque, dorm-room, drugstore, gymnasium (indoor), home-office, kindergarten-classroom, locker-room, mosque (outdoor), playground, playroom, pub (indoor), sandbox, schoolhouse, ski-resort, ski-slope, slum, swimming-hole, train-interior, train station/platform, tree-house, and waiting-room*. Note that we do not force the target class to be out of the 60 private classes, as this would restrict the diversity of target class and would disclose the private classes.

[3]https://github.com/CSAILVision/places365

B. Implementation details

We compare all FGSM-based attacks with the same parameters proposed in Kurakin *et al.* [12]. The adversarial noise per iteration is $\delta = 1$, which corresponds to the smallest variation in an 8-bit image. We constrain the maximum perturbation magnitude by setting $\epsilon = 16$, to provide a trade-off between the misleading rate and image quality degradation [13]. For a fair comparison, we ensure that all FGSM-based attacks, either employing one classifier or an ensemble of classifiers (E-FGSM and DI-FGSM), perform the same number of forward/backward passes on the classifiers, with $\overline{N} = \min(1.25\,\epsilon, \epsilon + 4)$ [12] iterations. The number of iterations for RP-FGSM is $N = \overline{N} \cdot K$, as only one classifier is used at each iteration (ensembles use all classifiers at each iteration). This ensures that all attacks that employ K classifiers perform $K \cdot \overline{N}$ forward/backward passes in total.

For EOT, we use as 2D transformations, $\mathbf{\Phi}_{2D}$, scaling, translation, rotation, lightening, darkening and additive Gaussian noise. The specific parameters of the transformations are chosen randomly from pre-defined intervals as follows: scaling with a factor between 0.8 and 1.2; translation between -0.2 and 0.2 times W; rotation between -60 and 60 degrees; lightening/darkening between 0 and 13 intensity points; and Gaussian noise with zero mean and 25 variance. We set $\lambda = 0.5$ to achieve a misleading rate of approximately 90%, as suggested in [5]. For DI-FGSM, we employ random resizing followed by random padding with a probability of 50% at each iteration. Original images $\mathbf{x} \in \mathbb{R}^{224\times224\times C}$ are resized and padded to $r \times r \times C$ with r randomly chosen within the range $[200, 224)$.

The parameters for the defense transformations are: for re-quantization, 1 to 7 bits per color channel in steps of 1 [7]; for median filtering, squared kernel of dimensions 2, 3 and 5 [7]; and for lossy JPEG compression, quality parameters 25, 50, 75 and 100 [24], [25]. For RP-FGSM, we randomly and uniformly select the transformation to apply to the image at each iteration, by selecting the parameters (i.e. number of bits for re-quantization, kernel dimension for median filtering and compression factor for JPEG compression) from the same sets used for the defense. As Li *et al.* [1], in targeted attacks, we use $\gamma = 0.99$ for target class selection.

C. Performance measures

We consider as performance measures the misleading rate, detectability and image quality.

Misleading rate is the ratio between the number of adversarial images that mislead the classifier with respect to the *true* class and the total number of images, expressed as a percentage. The higher the misleading rate in top ranks, the better the privacy protection. We consider as ranks top-1 and top-5. The misleading rate is calculated for seen and unseen classifiers, and with and without defenses.

Detectability is the percentage of correctly detected adversarial images with respect to the total number of images. We distinguish between adversarial images and original images by learning a threshold, τ, for each defense (re-quantization, median and JPEG compression) by accepting 5% of the

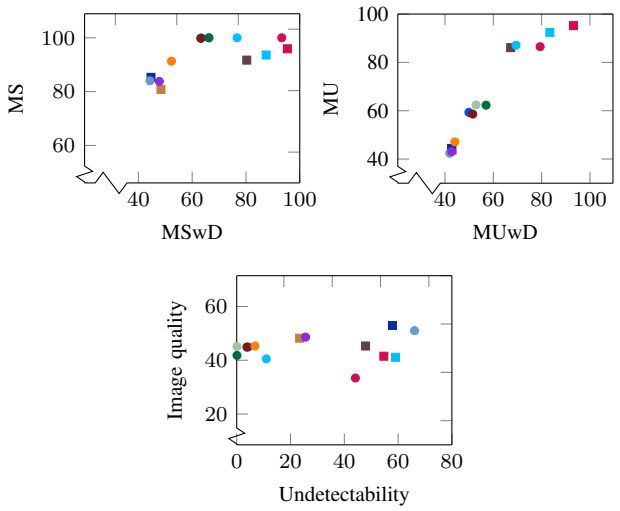


Fig. 4. Comparison of the privacy properties for: JSMA ●, CW ●, DeepFool ■, SparseFool ■, U-FGSM ■, R-FGSM ●, L-FGSM ●, P-FGSM ●, EOT ●, DI-FGSM (targeted) ●, DI-FGSM (untargeted) ■, E-FGSM ●, RP-FGSM (targeted) ● and RP-FGSM (untargeted) ■. For visualization purposes, all measures are normalized in the range 0-100%. KEY: MS, misleading a seen classifier; MU, misleading an unseen classifier; MSwD, MS with defense; MUwD, MU with defense; Image quality measured by BRISQUE [26].

original (training) images as adversarial (i.e. 5% false positive detection rate) (Eq. 4) [7].

Image quality is evaluated with Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) [26], a no-reference measure; and Peak-Signal-to-Noise Ratio (PSNR) and Most Apparent Distortion (MAD) [27], two full-reference measures. BRISQUE [26] quantifies distortions and unnaturalness. The lower the BRISQUE score, the better the image quality. PSNR quantifies the pixel-by-pixel difference between two images. The higher the PSNR, the smaller the added perturbation (and therefore, the better the image quality). MAD aims at measuring the perceived quality under different levels of distortion [27]. The lower the MAD score, the higher the image quality.

D. Evaluation of privacy protection

Figure 4 compares the results of RP-FGSM with those of the eleven considered state-of-the-art attacks. For visualization purposes, all metrics are normalized using their known ranges to 0% and 100%, where a higher value is more desirable. We report undetectability, the inverse of detectability; the top-1 misleading rate with seen (unseen) classifiers obtained with ResNet50 (DenseNet161); and image quality in terms of naturalness, BRISQUE. We report the results of the attacks when using the classifier (or combination of classifiers) with the highest misleading rate: ResNet50 for attacks that use a single classifier, and the combination of ResNet18, ResNet50 and AlexNet for attacks that use more than one classifier.

The top row of Figure 4 shows the trade-off between misleading seen (unseen) classifiers with and without defenses. The proposed attack outperforms the other attacks in these two plots (i.e. is nearer to the top-right corner of the plot). The second row of Figure 4 shows the relationship between image quality and undetectability. We can observe that all attacks obtain similar image quality, whereas attacks can be divided into two groups for the undetectability, those below 40% (CW, DeepFool, R-FGSM, L-FGSM, P-FGSM, EOT, E-FGSM, and DI-FGSM targeted), and those above 40% (JSMA, SparseFool, U-FGSM, and DI-FGSM untargeted). The proposed attack, RP-FGSM, in both targeted and untargeted versions, is within the group of less detectable attacks.

Table III summarizes all performance measures.

For *seen classifiers without defenses*, JSMA, CW, DeepFool and SparseFool have the lowest misleading rate (lower than 32.1% in top-5) as they craft their perturbation towards the closest class to the predicted class and the resulting class of the adversarial image could be the true class that we aim to protect (see Table II). Instead, targeted (untargeted) FGSM-based attacks have higher misleading rates because they move the prediction of the adversarial image closer to (farther from) the target (predicted) class at each iteration. The misleading rates of EOT are only between 67.4% and 72.7% in top-5, showing a limited ability to mislead the classifier when no transformation is applied. The misleading rates of untargeted FGSM-based attacks, such as U-FGSM and DI-FGSM, are above 83.4% in top-5, whereas those of RP-FGSM are above 86.9% when attacking a single classifier. Targeted FGSM-based attacks such as L-FGSM and P-FGSM obtain the highest misleading rates (above 95% in top-5). Similar results are obtained by attacks that use an ensemble of classifiers (E-FGSM and DI-FGSM). The target (untargeted) version of RP-FGSM obtains comparable results to the other FGSM-based attacks, i.e. above 97.8% (86.9%) in the top-5.

For an *unseen classifier*, JSMA, CW, DeepFool, SparseFool and EOT obtain low misleading rates (under 21.6% in the top-5). Similarly to the above results for seen classifiers, targeted FGSM attacks such as R-FGSM, L-FGSM, P-FGSM and E-FGSM (attacking a single classifier) obtain a higher performance, with misleading rates between 15.5% and 88.7% in the top-5. The highest misleading rates are provided by DI-FGSM and RP-FGSM when using three classifiers for generating the attack. In their targeted versions, they obtain misleading rates of 65.0% and 64.6% in the top-5, respectively, while the untargeted versions reach 76.0% and 82.6%, respectively. Adversarial attacks that combine classifiers obtain the highest misleading rates in unseen classifiers. Targeted DI-FGSM and RP-FGSM have comparable results (0.4 percentage points difference), while untargeted RP-FGSM outperforms untargeted DI-FGSM by more than 6 percentage points, thus indicating that the random selection of classifiers and transformations is an effective strategy for privacy protection.

For *seen classifiers with defenses*, JSMA, CW, DeepFool and SparseFool have low misleading rates (less than 21.2% in the top-5). Targeted attacks such as R-FGSM, L-FGSM, P-FGSM, EOT and E-FGSM (attacking one classifier) obtain much lower misleading rates compared to when no defenses are employed (19.8% - 53.4% in the top-5). Also in this case, untargeted FGSM-based attacks perform the best in the top-5 (above 47.4% for U-FGSM and above 59.3% for DI-FGSM). RP-FGSM outperforms all attacks when three classifiers are employed for crafting the adversarial images, with misleading

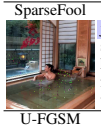
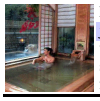
TABLE III

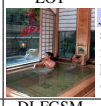
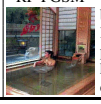
PRIVACY PROTECTION RESULTS IN THE TEST SPLIT OF PRIVATE PLACES365 DATASET FOR MISLEADING RATE, WITH AND WITHOUT DEFENSES, DETECTABILITY AND IMAGE QUALITY. MEASURES ARE REPORTED AS AVERAGE OF THE DATASET. STANDARD DEVIATION FOR IMAGE QUALITY IS SHOWN IN BRACKETS. GRAY SHADING INDICATES UNSEEN CLASSIFIER. BOLD FONT INDICATES THE BEST PERFORMING ATTACK FOR A GIVEN PERFORMANCE MEASURE. KEY – ATT.: ATTACK; R18: RESNET18; R50: RESNET50; A: ALEXNET; DN: DENSENET161; DET.: DETECTABILITY; BRISQUE: BLIND REFERENCELESS IMAGE SPATIAL QUALITY EVALUATOR; PSNR: PEAK-SIGNAL-TO-NOISE RATIO; MAD: MOST APPARENT DISTORTION; T1: TOP-1 MISLEADING RATE; T5: TOP-5 MISLEADING RATE; JSMA: JACOBIAN-BASED SALIENCY MAP ATTACK; CW: CARLINIWAGNER; DF: DEEPFOOL; SF: SPARSEFOOL; U-FG: UNTARGETED FGSM; R-FG: RANDOM FGSM; L-FG: LEAST-LIKELY FGSM; P-FG: PRIVATE FGSM; EOT: EXPECTATION OVER TRANSFORMATION; DI-FG: DIVERSE INPUT FGSM; E-FG: ENSEMBLE FGSM; RP-FGSM: ROBUST PRIVATE FGSM, THE PROPOSED ATTACK; T: TARGETED ATTACK; U: UNTARGETED ATTACK.

Att.	Classifier	Misleading ↑								Misleading with defenses ↑								Det. ↓	Image quality					
		R18		R50		A		DN		R18		R50		A		DN			BRISQUE ↓		PSNR ↑		MAD ↓	
		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5							
JSMA	R18	85.3	16.4	44.5	13.8	52.6	21.3	42.1	13.4	45.4	14.9	44.0	13.3	51.7	19.9	42.1	13.6	25.3	51.2	(13.9)	50.6 (7.4)		16.2	(8.1)
	R50	46.4	16.2	84.0	14.6	52.7	21.3	42.4	13.5	44.6	15.1	44.3	13.8	51.5	19.7	42.0	13.0	33.9	51.0	(14.0)	50.4 (7.4)		12.6	(7.3)
	A	45.6	15.6	43.8	13.8	85.6	22.3	41.5	13.3	44.7	15.0	43.5	13.3	52.2	19.9	41.5	13.3	26.8	51.0	(14.1)	48.8 (7.1)		13.0	(7.4)
CW	R18	83.8	25.6	44.7	13.8	52.3	21.3	42.5	13.6	50.6	16.4	44.0	13.3	51.5	19.9	42.4	13.5	72.0	48.5	(14.0)	**51.1** (0.4)		**0.0**	(0.1)
	R50	46.9	16.3	83.8	22.7	52.5	21.1	43.1	13.8	45.1	15.2	47.8	14.7	51.7	19.7	42.8	12.8	74.4	48.6	(14.0)	51.0 (0.4)		**0.0**	(0.1)
	A	45.2	15.8	43.6	13.7	84.4	30.0	41.7	13.5	44.6	15.0	43.3	13.4	56.1	20.9	41.8	13.4	71.9	48.8	(14.1)	51.0 (0.5)		**0.0**	(0.1)
DF	R18	82.6	28.5	47.0	16.6	45.2	15.9	42.8	13.6	50.5	16.7	43.7	13.3	51.6	20.0	42.6	13.4	74.6	48.1	(13.7)	50.1 (3.7)		1.4	(7.4)
	R50	44.9	13.6	80.8	24.5	43.3	13.8	43.7	13.9	45.6	15.2	48.4	15.4	51.8	19.8	43.0	13.0	76.6	48.2	(13.7)	49.6 (4.3)		2.1	(10.6)
	A	52.4	21.3	52.6	21.3	82.1	32.1	42.0	13.5	44.9	14.9	43.4	13.4	56.5	21.2	41.9	13.5	74.3	48.5	(13.7)	50.0 (3.5)		2.8	(12.2)
SF	R18	85.9	18.0	46.1	14.4	55.3	23.0	43.3	14.3	45.4	15.1	44.0	13.2	51.9	20.2	42.8	13.2	34.8	53.1	(13.9)	38.9 (8.6)		44.5	(31.7)
	R50	50.0	17.7	85.3	15.1	56.3	22.6	44.4	14.4	44.9	15.2	44.7	13.5	52.7	21.2	42.7	13.0	42.1	52.9	(13.9)	38.4 (8.7)		45.1	(32.0)
	A	46.4	16.3	44.2	14.0	85.9	23.0	41.7	13.6	45.5	15.1	43.7	13.5	53.6	20.4	41.8	13.3	**22.4**	51.4	(14.0)	40.7 (7.7)		28.5	(24.9)
U-FG	R18	92.4	86.1	86.6	55.9	67.8	34.3	85.5	53.5	86.1	59.7	67.0	32.6	62.5	29.4	67.1	33.4	58.1	45.6	(12.5)	32.0 (0.9)		26.1	(8.3)
	R50	85.9	54.4	91.7	84.3	65.0	31.0	86.1	53.3	69.6	34.1	80.3	47.4	61.5	27.0	67.2	32.4	52.1	45.3	(12.8)	32.5 (0.9)		24.1	(8.3)
	A	72.3	40.2	67.0	32.9	93.8	88.7	63.5	28.6	66.4	31.8	60.8	27.2	91.2	73.8	58.0	24.5	47.2	**29.7**	(17.1)	29.9 (0.8)		34.0	(9.2)
R-FG	R18	72.3	40.2	67.0	32.9	93.8	88.7	63.5	28.6	69.9	40.6	53.3	21.2	56.5	25.6	50.2	19.5	94.3	45.9	(12.7)	33.1 (1.0)		20.7	(8.3)
	R50	63.1	32.7	99.9	97.0	59.3	27.2	59.3	27.4	55.3	23.0	63.4	31.4	56.0	24.2	49.9	18.8	95.8	44.9	(13.0)	33.4 (1.0)		20.2	(8.3)
	A	53.9	23.1	52.3	19.9	99.8	96.4	49.6	17.8	52.1	20.0	50.1	19.0	74.0	46.5	47.6	17.0	96.7	34.7	(16.5)	32.4 (1.0)		19.3	(8.2)
L-FG	R18	**100.0**	**99.8**	64.4	34.4	62.0	30.1	61.5	32.0	74.1	45.3	54.5	23.4	58.4	27.1	52.1	21.3	99.3	46.3	(12.6)	32.7 (1.0)		19.2	(8.2)
	R50	68.0	38.8	**100.0**	**99.7**	60.3	28.1	62.4	32.7	57.0	25.0	66.3	34.6	57.2	25.2	52.9	22.2	99.9	45.1	(13.1)	32.9 (0.9)		18.7	(8.3)
	A	56.9	25.8	53.4	21.7	**100.0**	**99.7**	51.3	19.5	54.0	23.2	51.6	19.9	79.0	53.4	49.5	18.4	99.6	34.9	(16.5)	32.0 (0.8)		19.6	(8.3)
P-FG	R18	**100.0**	97.4	60.4	29.0	60.7	28.3	57.0	26.9	71.2	39.1	53.5	20.7	57.3	25.4	51.3	19.2	93.6	45.9	(12.7)	33.2 (1.0)		22.4	(8.4)
	R50	63.8	33.0	99.8	97.0	59.5	26.9	58.6	27.3	55.0	23.1	63.2	30.7	56.6	24.7	51.5	19.3	96.2	44.9	(13.0)	33.3 (1.0)		21.4	(8.5)
	A	55.1	23.4	52.9	19.7	99.7	96.1	49.6	18.5	52.5	20.6	50.2	19.0	75.1	47.0	47.8	17.4	95.4	34.8	(16.4)	32.4 (0.9)		24.4	(9.0)
EOT	R18	94.0	72.0	48.3	17.5	54.2	22.5	46.1	15.8	56.8	23.0	46.9	15.5	52.8	21.6	44.0	14.4	93.4	45.3	(13.6)	42.3 (2.1)		3.2	(4.0)
	R50	50.9	20.6	91.3	67.4	54.0	22.2	47.1	16.9	48.6	17.9	52.3	19.8	52.9	21.2	44.1	14.8	93.2	45.3	(13.6)	42.8 (1.9)		3.4	(4.2)
	A	46.9	16.9	45.3	14.2	90.8	72.7	42.3	13.7	45.2	15.3	44.3	13.7	62.2	31.6	42.3	13.1	93.9	43.8	(14.4)	43.3 (2.0)		3.6	(4.3)
DI-FG (T)	R18	**100.0**	98.9	71.1	43.1	71.5	42.3	68.4	39.6	82.7	56.2	61.5	30.1	60.1	29.2	58.2	26.9	82.8	43.6	(13.5)	32.8 (0.7)		26.2	(8.8)
	R50	80.8	54.5	**100.0**	98.3	64.5	32.6	76.6	49.6	66.4	36.1	74.8	44.7	59.6	28.5	61.8	30.5	86.7	43.6	(13.5)	33.0 (0.7)		25.3	(8.9)
	A	60.5	28.0	54.8	23.4	**100.0**	97.9	52.5	20.4	56.9	25.7	53.9	22.1	85.2	64.0	50.9	18.8	86.6	36.6	(16.1)	32.6 (0.6)		26.5	(9.1)
	R18+R50	**100.0**	99.1	**100.0**	98.8	68.4	37.1	86.5	62.8	83.8	59.9	77.0	49.2	63.2	31.8	68.7	38.9	85.0	43.5	(13.5)	32.7 (0.7)		27.6	(8.9)
	R18+A	**100.0**	98.6	71.5	42.3	**100.0**	98.3	67.4	37.8	79.9	54.9	63.1	31.8	83.6	62.0	58.9	28.7	87.8	39.6	(15.0)	32.2 (0.6)		28.2	(9.0)
	R50+A	83.2	59.6	**100.0**	98.2	**100.0**	98.4	75.2	48.1	71.4	42.4	74.1	45.2	82.8	61.4	62.8	32.5	92.2	39.8	(14.9)	32.3 (0.6)		27.5	(9.1)
	R18+R50+A	**100.0**	98.5	**100.0**	98.4	99.9	98.2	87.1	65.0	83.2	59.0	76.7	48.5	83.3	61.6	69.4	40.1	89.0	40.5	(14.7)	32.0 (0.7)		29.1	(9.1)
DI-FG (U)	R18	92.4	86.0	89.5	67.6	73.5	41.3	89.1	64.6	89.7	71.7	74.6	43.4	67.6	34.8	73.9	42.1	53.2	43.4	(13.5)	32.9 (0.7)		25.8	(8.5)
	R50	89.7	68.2	91.8	84.5	70.6	37.7	90.0	67.0	77.5	46.6	87.0	61.5	65.9	31.3	76.7	42.0	46.6	43.8	(13.5)	33.2 (0.7)		24.4	(8.5)
	A	74.3	42.8	66.5	32.7	94.3	89.6	62.6	28.3	70.2	36.7	64.0	29.9	92.9	78.5	59.4	26.4	46.8	34.9	(16.6)	32.2 (0.6)		27.5	(9.0)
	R18+R50	93.2	87.2	93.1	86.2	78.0	46.3	92.2	76.1	90.4	72.9	88.0	64.0	72.0	37.7	81.7	51.5	42.7	43.5	(13.5)	32.7 (0.7)		26.4	(8.6)
	R18+A	92.2	86.1	89.6	67.3	94.0	87.8	88.4	64.4	89.4	70.5	76.8	46.1	90.3	72.5	76.6	44.5	51.1	40.0	(14.4)	32.3 (0.7)		27.6	(8.7)
	R50+A	88.9	68.3	91.2	83.4	93.6	87.6	87.9	64.7	79.8	50.6	84.7	59.3	90.2	72.1	76.8	43.7	42.0	40.0	(14.4)	32.5 (0.7)		26.4	(8.6)
	R18+R50+A	93.6	87.4	93.6	86.4	94.3	87.4	92.4	76.0	90.7	72.9	87.5	64.3	89.1	70.4	83.4	53.7	41.0	41.1	(14.0)	32.3 (0.7)		27.6	(8.7)
E-FG	R18	**100.0**	96.4	51.9	20.2	56.2	23.7	48.1	18.6	64.5	32.8	48.5	18.0	54.1	22.1	46.2	15.8	99.9	45.3	(13.3)	37.5 (2.3)		9.4	(7.5)
	R50	55.1	24.9	**100.0**	95.9	55.8	23.8	50.7	19.7	51.2	20.2	58.8	27.1	53.8	21.9	46.0	16.2	99.9	44.7	(13.4)	37.3 (2.3)		9.7	(7.6)
	A	49.2	19.4	48.5	16.3	**100.0**	97.0	45.5	15.5	48.9	17.6	47.4	15.6	69.2	39.7	44.9	14.5	99.8	38.9	(15.9)	36.1 (2.3)		13.3	(8.7)
	R18+R50	**100.0**	97.4	**100.0**	96.9	58.1	26.1	58.7	27.3	67.0	35.6	60.9	29.0	55.1	23.9	50.1	19.9	100.0	44.9	(13.3)	35.7 (2.4)		13.9	(9.1)
	R18+A	**100.0**	96.5	55.1	24.0	**100.0**	96.7	52.7	21.8	66.6	36.5	50.9	19.5	70.6	42.1	48.9	18.1	99.8	41.1	(14.7)	34.6 (2.3)		17.5	(9.8)
	R50+A	59.6	28.9	**100.0**	96.1	**100.0**	96.6	54.7	22.6	54.4	22.3	59.6	28.2	70.1	41.3	49.4	18.4	99.9	40.9	(14.7)	34.6 (2.2)		17.2	(9.7)
	R18+R50+A	**100.0**	97.4	**100.0**	96.8	**100.0**	97.3	62.3	31.4	73.5	43.8	66.2	35.2	73.9	46.5	57.0	24.7	99.9	41.8	(14.2)	33.8 (2.3)		19.8	(9.6)
RP-FG (T)	R18	**100.0**	98.4	62.6	31.8	62.3	29.7	60.4	30.1	94.4	80.1	58.6	27.9	62.0	30.6	56.0	25.1	67.0	44.8	(13.2)	33.6 (0.6)		23.0	(9.0)
	R50	68.2	38.6	**100.0**	97.8	61.0	28.9	62.8	31.8	62.5	31.5	85.7	59.6	61.0	29.1	57.4	27.6	79.0	44.2	(13.4)	33.8 (0.5)		22.2	(9.0)
	A	54.6	23.8	51.5	20.4	**100.0**	98.1	49.4	18.3	53.0	22.6	50.8	20.0	91.0	74.2	48.6	18.0	64.1	37.7	(15.7)	33.2 (0.6)		23.4	(9.1)
	R18+R50	**100.0**	98.9	**100.0**	98.7	68.2	39.1	84.4	59.6	96.5	85.5	93.7	80.1	65.4	35.0	75.7	49.0	58.6	30.6	(7.0)	31.0 (0.5)		34.0	(10.9)
	R18+A	**100.0**	98.7	71.4	43.5	**100.0**	98.2	67.2	39.8	93.8	78.2	66.5	36.5	91.9	75.8	63.0	33.3	65.5	30.6	(7.6)	30.8 (0.5)		33.9	(10.7)
	R50+A	79.9	53.2	99.9	98.6	**100.0**	98.6	70.5	42.3	72.8	44.6	90.3	72.4	91.7	75.8	65.8	35.8	72.9	30.1	(7.5)	30.8 (0.5)		33.6	(10.8)
	R18+R50+A	**100.0**	99.1	**100.0**	98.6	**100.0**	98.2	86.5	64.6	**96.8**	88.0	93.3	79.3	91.7	77.7	79.4	55.1	55.9	33.4	(7.4)	29.6 (0.4)		39.8	(11.5)
RP-FG (U)	R18	93.5	88.5	88.3	62.6	70.1	37.7	86.4	59.0	93.2	83.9	80.7	53.3	67.1	34.3	80.9	50.0	50.6	44.6	(13.1)	33.3 (0.6)		23.7	(8.8)
	R50	88.3	63.1	92.9	86.9	67.6	34.9	88.1	60.5	83.6	55.2	92.3	79.3	65.4	32.1	81.7	52.7	42.9	44.2	(13.4)	33.6 (0.6)		22.5	(8.8)
	A	67.6	34.4	60.1	26.7	95.6	91.1	56.8	22.8	66.0	33.1	59.6	26.1	**94.9**	86.6	56.7	22.7	47.3	36.1	(15.9)	32.4 (0.5)		26.1	(9.2)
	R18+R50	95.1	90.6	94.4	88.6	79.6	50.2	93.5	78.5	93.6	83.3	93.8	84.5	76.2	45.9	89.8	67.5	41.7	45.3	(12.5)	30.9 (0.5)		34.1	(10.8)
	R18+A	94.7	89.6	90.9	69.3	95.3	90.3	89.2	66.2	92.8	79.3	81.4	52.8	**94.9**	88.1	80.8	51.7	57.6	40.4	(13.9)	30.5 (0.5)		34.8	(10.6)
	R50+A	91.6	72.5	94.0	87.8	95.2	90.3	90.6	68.0	85.4	59.8	90.1	70.3	94.7	**88.4**	82.0	50.7	49.0	40.0	(13.9)	30.6 (0.5)		34.2	(10.6)
	R18+R50+A	97.0	93.8	95.9	91.0	95.7	91.0	**95.2**	**82.6**	96.5	**91.7**	**95.4**	**87.7**	**94.9**	84.4	**93.2**	**77.0**	45.3	41.5	(13.1)	29.4 (0.4)		40.5	(11.3)

TABLE IV

EXAMPLE OF ADVERSARIAL ATTACKS AND TOP-5 PREDICTIONS WITH DIFFERENT CLASSIFIERS. ATTACKS SHOWN ARE THE BEST-PERFORMING WITH RESPECT TO THE SET OF PRIVACY-PRESERVING PERFORMANCE MEASURES. CW, DEEPFOOL, SPARSEFOOL, U-FGSM, P-FGSM, AND EOT USE RESNET50; DI-FGSM, E-FGSM AND RP-FGSM USE A COMBINATION OF RESNET18, RESNET50 AND ALEXNET; DI-FGSM AND RP-FGSM ARE PRESENTED WITH THEIR UNTARGETED VERSIONS. FIRST AND SIXTH COLUMN: ORIGINAL OR ADVERSARIAL IMAGE; OTHER COLUMNS: TOP-5 PREDICTED CLASSES AND CORRESPONDING PROBABILITY VALUE; BLUE SHADING: TRUE CLASS; $\triangle$: PROBABILITY HIGHER THAN 99.95; $\triangledown$: PROBABILITY LOWER THAN 0.05.

	ResNet18		ResNet50		AlexNet		DenseNet161	
Original								
	jacuzzi	87.2	jacuzzi	87.4	jacuzzi	83.5	jacuzzi	90.4
	swim pool ind.	7.1	swim pool ind.	6.7	swim pool ind.	6.7	hot spring	6.4
	hot spring	2.5	hot spring	3.5	patio	1.5	swim pool ind.	3.0
	fountain	0.9	sauna	0.6	hot spring	1.2	water park	0.1
	water park	0.7	water park	0.5	porch	1.1	sauna	$\triangledown$0.0
CW								
	swim pool ind.	52.0	swim pool ind.	52.0	jacuzzi	82.3	jacuzzi	86.8
	jacuzzi	31.6	jacuzzi	31.6	swim pool ind.	7.2	hot spring	7.8
	hot spring	4.7	hot spring	4.7	patio	1.5	swim pool ind.	5.0
	water park	3.4	water park	3.4	hot spring	1.2	water park	0.1
	swim pool out.	2.7	swim pool out.	2.7	porch	1.1	sauna	0.1
DeepFool								
	jacuzzi	82.9	hot spring	64.4	jacuzzi	82.7	jacuzzi	83.9
	swim pool ind.	8.9	jacuzzi	24.0	swim pool ind.	6.8	hot spring	12.3
	hot spring	3.7	swim pool ind.	4.8	patio	1.6	swim pool ind.	3.3
	water park	1.3	water park	1.6	hot spring	1.3	water park	0.1
	fountain	1.1	sauna	1.0	porch	1.1	sauna	0.1
SparseFool								
	jacuzzi	88.5	hot spring	46.1	jacuzzi	80.9	jacuzzi	85.6
	swim pool ind.	6.6	jacuzzi	45.7	swim pool ind.	6.5	hot spring	11.0
	hot spring	2.3	swim pool ind.	5.2	patio	2.1	swim pool ind.	3.0
	fountain	0.9	sauna	0.9	fishpond	1.6	water park	0.1
	water park	0.8	water park	0.5	hot spring	1.3	sauna	0.1
U-FGSM								
	hot spring	88.3	hot spring	$\triangle$100.0	jacuzzi	83.5	hot spring	95.8
	jacuzzi	5.2	volcano	$\triangledown$0.0	swim pool ind.	6.7	jacuzzi	3.3
	water park	3.1	fountain	$\triangledown$0.0	patio	1.5	swim pool ind.	0.5
	swim pool ind.	1.3	waterfall	$\triangledown$0.0	hot spring	1.2	fountain	0.2
	fountain	0.9	mountain	$\triangledown$0.0	porch	1.1	water park	0.1

	ResNet18		ResNet50		AlexNet		DenseNet161	
P-FGSM								
	jacuzzi	77.3	garage ind.	$\triangle$100.0	jacuzzi	62.2	jacuzzi	77.2
	swim pool ind.	19.3	parking	$\triangledown$0.0	fishpond	6.2	swim pool ind.	19.3
	fountain	0.5	garage out.	$\triangledown$0.0	patio	5.4	lobby	0.6
	hot spring	0.5	bus station	$\triangledown$0.0	swim pool ind.	4.6	fountain	0.4
	water park	0.3	subway st.	$\triangledown$0.0	loading dock	2.6	hot spring	0.2
EOT								
	jacuzzi	84.1	greenho. ind.	52.5	jacuzzi	81.8	jacuzzi	93.6
	swim pool ind.	4.9	greenho. out.	44.9	swim pool ind.	5.1	swim pool ind.	2.7
	fountain	3.7	pet shop	1.8	fishpond	3.2	hot spring	1.8
	hot spring	1.7	roof garden	0.5	patio	1.7	fountain	0.3
	water park	1.7	aquarium	0.1	hot spring	1.3	fishpond	0.3
DI-FGSM								
	amus. park	98.9	fastfood rest.	91.7	lock chamb	56.7	rest. patio	17.6
	carrousel	0.9	food court	3.2	canal	21.9	diner	14.7
	water park	0.2	icecream parlor	1.1	boathouse	14.9	food court	13.4
	playground	$\triangledown$0.0	coffee shop	0.9	lake	1.4	amus. park	8.9
	ticket booth	$\triangledown$0.0	cafeteria	0.8	bridge	1.1	coffee shop	8.5
E-FGSM								
	foot. field	99.5	foot. field	99.0	foot. field	99.7	jacuzzi	74.6
	stadium	0.4	stadium foot.	0.9	stadium	0.2	swim pool ind.	6.4
	soc. field	$\triangledown$0.0	stadium base.	$\triangledown$0.0	soc. field	0.1	hot spring	5.6
	athl. field	$\triangledown$0.0	athl. field	$\triangledown$0.0	basket. court	$\triangledown$0.0	fountain	5.2
	stadium	$\triangledown$0.0	soc. field	$\triangledown$0.0	stadium	$\triangledown$0.0	water park	1.6
RP-FGSM								
	boathouse	96.0	fastfood rest.	40.8	lock chamb	99.9	amus. park	27.3
	lock chamb	3.3	icecream parlor	12.7	boathouse	$\triangledown$0.0	boathouse	9.7
	pier	0.2	coffee shop	12.6	bridge	$\triangledown$0.0	lock chamb	7.2
	harbor	0.2	pizzeria	12.4	canal urban	$\triangledown$0.0	pet shop	4.0
	canal	0.1	diner	9.7	canal natural	$\triangledown$0.0	rail. track	3.7

rates in the top-5 between 59.6% and 88.0% when targeted, and between 70.3% and 91.7% when untargeted. When attacking three classifiers, the misleading rate of E-FGSM drops by 61.6 percentage points (35.2% from 96.8%) when defenses are applied. Instead, RP-FGSM untargeted only drops by 3.3 percentage points (87.7% from 91.0%).

When evaluating *misleading unseen classifiers with defenses*, JSMA, CW, DeepFool and SparseFool still obtain the lowest misleading rates (lower than 21.2% in top-5). Similarly, FGSM-based attacks such as R-FGSM, L-FGSM, P-FGSM, EOT and E-FGSM obtain very low misleading rates (under 27.1% in the top-5), while DI-FGSM, when attacking three classifiers, obtains higher misleading rates with values of 40.1% (targeted) and 53.7% (untargeted). Also in this case, the proposed RP-FGSM outperforms all other attacks, obtaining a misleading rate in the top-5 of 55.1% (targeted) and of 77.0% (untargeted). This confirms that untargeted attacks exhibit a higher degree of privacy protection with unseen classifiers, both with and without defenses.

Table IV compares predicted probabilities of sample adversarial images. We observe that most of the adversarial attacks are unsuitable for privacy protection as the true class might be within the top-5. DI-FGSM and RP-FGSM are the only attacks capable of removing the true class from the top-5 for all classifiers in both seen (ResNet18, ResNet50 and AlexNet) and unseen (DenseNet161) settings. This is mainly due to the fact that the attacks do not overfit the perturbations to a specific classifier and that the crafted perturbations generate adversarial predicted classes that are far from the predicted class in the decision space.

Targeted attacks are in general more detectable than their untargeted versions, as reaching the target class often means crafting a larger perturbation. R-FGSM, L-FGSM, P-FGSM, EOT and E-FGSM are highly *detectable*, with values above 93.2%, while U-FGSM obtains detection rates below 58.1%. The detectability of targeted DI-FGSM is between 82.8% and 92.2%, and that of targeted RP-FGSM between 55.9% and 79.0% (RP-FGSM is less detectable than DI-FGSM in this setting). Untargeted DI-FGSM obtains detectability between 41.0% and 53.2%, while untargeted RP-FGSM between 41.7% and 57.6%, thus DI-FGSM is slightly less detectable than RP-FGSM. JSMA and SparseFool have the lowest detection rates, with values under 42.1%.

Attacks with a sparse perturbation, such as JSMA and SparseFool, have the highest *image quality* results (PSNR values above 38.4 dB). However, sparse perturbations are often of high magnitude (Figure 1) and, therefore, highly noticeable (Table IV). This effect is reflected by the poor BRISQUE score, above 51.0 points, for JSMA and SparseFool. Attacks with dense perturbations obtain a lower PSNR than sparse ones, whose accumulated variation of intensity across the whole image is smaller. However, dense perturbations might be more desirable for privacy protection, as they are less visible and their per-pixel intensity variation is lower than that of sparse attacks. FGSM-based attacks obtain consistently better BRISQUE scores than non-FGSM based attacks. FGSM-based attacks obtain around 44 BRISQUE score and non-FGSM based attacks obtain around 5 points more (the lower the better). EOT, which has an image quality regularization in its modeling (see Eq. 12), is the FGSM-based attack with the highest PSNR results (above 42.3 dB) and the lowest MAD score (below 3.6 points). However, its BRISQUE scores are higher (i.e. worse) than those of other FGSM-based attacks, thus indicating a lower image naturalness. When the number

TABLE V

IMPACT ON MISLEADING AND DETECTABILITY OF USING A RANDOM TRANSFORMATION IN CRAFTING ADVERSARIAL PERTURBATIONS (VALUES ARE THE VARIATION OF THE SCORE COMPARED TO THE SCORE OBTAINED WHEN NO TRANSFORMATION IS USED). GREEN SHADING REPRESENTS AN INCREASE IN MISLEADING RATE OR A DECREASE IN DETECTABILITY. RED SHADING REPRESENTS A DECREASE IN MISLEADING RATE OR AN INCREASE IN DETECTABILITY. KEY– CLASS. COMB.: CLASSIFIER COMBINATION; N/A: NOT APPLICABLE; R18: RESNET18; R50: RESNET50; A: ALEXNET; DN: DENSENET161; T1: TOP-1 MISLEADING RATE; T5: TOP-5 MISLEADING RATE.

		Targeted																	Untargeted													
Class. comb.	Classifier	Misleading ↑								Misleading with defense ↑								Det. ↓	Misleading ↑								Misleading with defense ↑								Det. ↓
		R18		R50		A		DN		R18		R50		A		DN			R18		R50		A		DN		R18		R50		A		DN		
		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	
N/A	R18	0.0	0.1	2.6	2.7	1.4	2.6	2.1	2.6	-0.3	-3.3	-0.7	-2.2	-2.1	-0.5	-1.3	-0.9	-23.0	1.9	2.2	1.5	6.8	2.5	3.5	1.1	5.6	26.2	52.7	-5.3	23.4	4.9	4.9	13.7	16.1	-7.4
	R50	4.6	5.6	0.0	0.1	2.5	2.1	4.2	4.2	6.6	5.7	22.5	30.2	2.1	3.3	7.6	6.7	-17.8	2.1	9.0	1.3	2.4	2.6	4.1	2.2	8.3	10.9	19.0	11.8	36.4	4.2	5.0	14.5	20.3	-8.8
	A	1.1	0.5	-0.2	0.1	0.0	0.0	-0.3	-0.1	1.8	2.1	1.2	0.8	14.7	24.4	1.1	0.2	-17.3	-4.9	-5.8	-7.0	-6.2	1.8	2.3	-6.5	-5.9	-0.4	1.5	-1.0	-1.4	3.5	13.0	-1.5	-1.4	-0.2
	Average	1.9	2.1	0.8	1.0	1.3	1.6	2.0	2.2	2.7	1.5	7.7	9.6	4.9	9.1	2.5	2.0	-19.4	-0.3	1.8	-1.4	1.0	2.3	3.3	-1.1	2.7	12.2	24.4	1.8	19.5	4.2	7.6	8.9	11.7	-5.5
Ensemble	R18+R50	0.0	-0.4	0.0	-0.7	1.7	2.1	3.4	4.9	18.6	29.2	10.5	16.5	3.3	2.4	10.3	12.7	-22.7	0.6	1.3	0.8	1.5	2.3	5.8	0.4	1.3	2.9	12.8	5.0	19.0	5.6	10.4	7.9	18.1	9.0
	R18+A	0.0	0.0	1.1	2.6	0.0	-0.6	1.5	2.3	19.1	28.5	5.6	6.2	8.5	18.2	28.1	-0.8	-18.7	1.0	1.7	0.5	3.4	0.9	2.5	0.5	2.7	3.6	14.0	8.9	15.1	3.7	12.0	8.6	15.2	8.7
	R50+A	4.2	5.3	-0.1	-0.9	0.0	-1.0	1.9	3.5	8.8	10.0	19.1	26.2	11.5	17.5	7.5	-18.8	-8.8	2.2	4.3	1.3	2.3	1.0	2.5	1.5	2.8	8.0	14.8	7.0	20.4	3.9	12.0	8.9	15.2	10.1
	R18+R50+A	0.0	-0.7	-0.1	-1.0	-0.1	-1.7	3.1	3.6	24.4	35.9	12.0	18.4	9.4	13.7	10.3	13.9	-12.6	1.9	4.3	2.2	4.6	1.6	5.2	2.8	8.1	4.7	17.7	8.1	24.9	6.6	17.8	10.9	27.5	9.8
	Average	1.1	1.1	0.2	0.0	0.4	-0.3	2.5	3.6	17.7	25.9	11.8	16.8	8.2	13.0	14.1	1.8	-15.7	1.7	3.5	1.4	3.4	1.2	3.4	1.6	4.5	5.4	15.5	8.0	20.1	4.7	13.9	9.5	19.3	9.5
Random	R18+R50	-0.4	-2.1	-0.2	-0.8	0.4	0.7	-1.5	-0.7	9.8	16.5	14.9	18.7	1.5	1.8	6.6	5.7	-14.0	0.9	1.8	1.0	1.6	4.0	6.9	1.3	6.8	9.0	27.4	4.6	17.8	6.1	9.5	11.9	24.1	-9.9
	R18+A	-0.6	-3.8	-1.9	-2.0	-0.2	0.3	-1.5	-0.9	8.9	11.0	2.6	2.5	11.5	17.6	2.4	1.9	-5.8	1.1	2.0	2.1	5.1	1.2	1.1	1.9	5.3	4.9	15.3	7.3	11.2	4.3	13.2	7.2	11.9	-2.8
	R50+A	-7.0	-8.2	-2.8	-9.4	-0.0	0.1	-5.8	-16.1	3.5	5.5	7.4	7.7	11.7	17.9	2.5	2.9	0.1	2.5	6.9	1.1	1.6	1.0	0.8	2.8	8.2	7.7	12.9	8.2	18.4	3.8	15.8	8.3	11.6	-2.9
	R18+R50+A	-0.3	-2.3	-2.3	-8.3	-1.3	-2.3	-2.9	-2.9	1.9	3.7	19.3	27.8	9.5	11.3	4.0	5.0	-7.9	2.4	4.5	2.1	3.8	-0.4	-0.6	2.5	11.4	5.1	27.1	8.8	31.6	3.1	13.9	12.4	30.7	-3.1
	Average	-2.1	-4.1	-1.8	-5.1	-0.3	-0.3	-2.9	-5.2	6.0	9.2	11.1	14.2	8.6	12.2	3.9	3.9	-6.9	2.0	4.5	1.8	3.5	0.6	0.4	2.4	8.3	5.9	18.4	8.1	20.4	3.7	14.3	9.3	18.1	-2.9

of classifiers used to create the adversarial images increases, a higher level of perturbations is required, which results in a decrease of image quality. For instance, we see a decrease of about 3-4 dBs when DI-FGSM, E-FGSM and RP-FGSM employ three classifiers instead of only one. However, FGSM-based attacks can obtain the desired trade-off between misleading rate and image quality by tuning the parameter ϵ (Eq. 11), a desirable feature for applications such as privacy protection (see the sensitivity analysis in Sec. V-E). RP-FGSM obtains quality scores comparable to other FGSM-based attacks in terms of BRISQUE and PSNR. For instance, when attacking three classifiers, E-FGSM obtains BRISQUE score 41.8 and PSNR 33.8 dB, whereas RP-FGSM obtains BRISQUE score 33.4 (the lower the better) and PSNR 29.6 dB (the higher the better). However, among the FGSM methods for three classifiers, E-FGSM obtains the best MAD score 19.8, whereas RP-FGSM obtains the worst MAD score 40.5 (the lower the better).

E. Ablation and sensitivity analysis

We present two analyses of RP-FGSM, namely an ablation study to validate the effectiveness of the random selection of classifier and transformation, and of different number of iterations; and a sensitivity analysis on the parameter ϵ.

The ablation study compares the effect of using a randomly selected transformation with respect to not using any transformations, and the effect of using randomly selected classifiers with respect to using an ensemble of classifiers at each iteration [6]. Furthermore, we evaluate the effect of the number of iterations, N, when attacking multiple classifiers. For a fair comparison, all these experiments follow the same sequence of transformations, classifiers and target class, whenever applicable, by fixing the random seed. We report the variation of performance measures, namely misleading rates with and without defense, and detectability, for both targeted and untargeted RP-FGSM.

Table V shows the effect of using random transformations with respect to not using any transformation. When employing a single classifier, using transformations generally improves the misleading and decreases the detectability. In particular, the detectability decreases by an average of 19.4 (5.5) percentage points in the targeted (untargeted) version. When using an ensemble of classifiers, the misleading generally increases for both targeted and untargeted. For the untargeted version, the detectability increases by an average of 9.5 percentage points, whereas for the targeted one, increases by an average of 15.7 percentage points. We argue that using an ensemble of classifiers generates overfitted perturbations that successfully mislead the classifier, but are more detectable. When using a random classifier at each iteration, as we propose, the misleading with defense increases in both targeted and untargeted versions. Moreover, the detectability in the untargeted attack decreases, contrary to when using an ensemble of classifiers. This experiment shows that introducing random transformations improves the performance, especially when a random classifier is employed at each iteration.

Table VI compares the effect of using a randomly selected classifier with respect to using an ensemble of classifiers, at each iteration. We report the results when not using any transformation, for reference, and also with a random selection of transformations. When random transformations are employed in the generation, the use of a randomly selected classifier at each iteration increases, in general, the misleading rate in the targeted version by a maximum of 4.2 percentage points in top-1 with an unseen classifier (DenseNet) . Also, detectability decreases by 12.5 (5.2) percentage points in the targeted (untargeted) version. Some instances of misleading rate decreases are observed, especially in the untargeted version.

Table VII reports how results vary using a random selection of the classifier with respect to an ensemble of classifiers, with varying numbers of iterations. We report the performance measures with the proposed number of iterations (i.e. 40, with two classifiers; and 60, with three classifiers), which allows the proposed attack to perform the same number of forward and backward passes on the classifiers compared to an ensemble FGSM (e.g. E-FGSM). We also study the effect of the number of iterations on the performance measures by considering fewer iterations (i.e. 20). For an equal number of

TABLE VI

IMPACT ON MISLEADING AND DETECTABILITY OF USING A RANDOMLY SELECTED CLASSIFIER IN CRAFTING ADVERSARIAL PERTURBATIONS (VALUES ARE THE VARIATION OF THE SCORE COMPARED TO THE SCORE OBTAINED WHEN COMBINING CLASSIFIERS AS AN ENSEMBLE [6]). GREEN SHADING REPRESENTS AN INCREASE IN MISLEADING RATE OR A DECREASE IN DETECTABILITY. RED SHADING REPRESENTS A DECREASE IN MISLEADING RATE OR AN INCREASE IN DETECTABILITY. KEY – TRANS.: TRANSFORMATION; N/A: NOT APPLICABLE; R18: RESNET18; R50: RESNET50; A: ALEXNET; DN: DENSENET161; T1: TOP-1 MISLEADING RATE; T5: TOP-5 MISLEADING RATE.

		Targeted																	Untargeted													
		Misleading ↑								Misleading with defense ↑									Misleading ↑								Misleading with defense ↑							
Trans.	Classifier	R18		R50		A		DN		R18		R50		A		DN		Det. ↓	R18		R50		A		DN		R18		R50		A		DN		Det. ↓
		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	
N/A	R18+R50	0.0	-0.2	0.0	-0.8	-0.5	-0.7	-2.0	-2.4	-1.2	-2.0	-10.1	-10.8	-0.3	-1.0	-1.6	-0.9	0.6	1.1	1.6	0.4	0.8	-2.5	-3.0	-0.1	-4.3	-5.8	-17.2	1.2	2.7	-1.8	-1.0	-3.9	-8.0	8.9
	R18+A	0.0	0.0	-2.1	-1.1	0.0	-0.6	-1.7	-1.5	-2.1	-1.9	-1.1	-1.3	-1.8	-1.9	21.9	-7.5	0.7	1.3	1.6	-0.8	-2.9	0.2	1.2	-1.0	-3.4	-1.5	-6.6	-2.6	-4.5	0.4	0.0	-2.8	-4.8	9.8
	R50+A	4.2	5.3	0.0	-1.1	-0.1	-0.8	1.9	13.0	-1.5	-1.2	-1.5	-1.2	-1.3	-2.5	-1.5	-26.9	-8.0	0.4	-2.6	1.7	2.8	0.6	1.7	0.2	-4.8	-2.1	-4.0	-2.8	-7.4	0.7	0.2	-3.4	-4.4	9.8
	R18+R50+A	0.0	-0.2	-0.1	-0.9	-0.2	-3.1	-2.4	-3.7	4.5	5.0	-8.1	-10.1	-2.7	-3.8	-2.6	-2.2	0.2	1.2	2.1	0.3	0.8	1.8	4.3	0.2	-4.4	0.7	-8.2	-1.0	-10.1	2.8	0.2	-2.5	-7.4	7.0
	Average	1.1	1.2	-0.5	-1.0	-0.2	-1.3	-1.1	1.4	-0.1	0.0	-5.2	-5.9	-1.5	-2.3	4.1	-9.4	-1.6	1.0	0.7	0.4	0.4	0.0	1.1	-0.2	-4.2	-2.2	-9.0	-1.3	-4.8	0.5	-0.2	-3.2	-6.2	8.9
Random	R18+R50	0.0	0.7	0.0	0.7	36.6	65.4	8.7	11.1	0.6	1.8	7.4	11.0	1.3	5.4	9.7	11.6	-7.9	1.3	2.1	0.6	0.9	-0.8	-1.9	0.8	1.2	0.3	-2.6	0.8	1.5	-1.3	-1.9	0.1	-2.0	-10.0
	R18+A	0.0	0.6	7.4	8.6	0.0	0.9	7.2	8.8	-3.3	-5.1	5.6	6.2	12.2	19.0	5.9	6.8	-3.4	1.4	1.9	0.8	-1.3	0.5	-0.2	0.4	-0.8	-0.2	-5.3	-4.2	-8.4	1.0	1.2	-4.2	-8.1	-1.7
	R50+A	6.7	10.0	0.1	1.4	0.0	0.5	8.7	10.1	5.8	8.3	-3.3	-5.0	8.9	18.8	5.1	7.2	-9.8	0.6	0.0	1.5	2.2	0.5	0.1	1.5	0.6	-2.4	-5.9	-1.6	-9.4	0.6	4.0	-4.0	-8.0	-3.2
	R18+R50+A	0.0	1.0	0.1	1.6	0.0	1.1	11.4	16.4	3.9	9.6	8.5	19.1	6.2	12.0	8.9	12.4	-28.7	1.7	2.2	0.1	0.0	-0.2	-1.5	0.0	-1.1	1.1	1.2	-0.3	-3.4	-0.7	-3.7	-1.0	-4.2	-5.9
	Average	1.7	3.1	1.9	3.1	9.2	17.0	9.0	11.6	1.8	3.7	4.6	7.8	7.2	13.8	7.4	9.5	-12.5	1.3	1.5	0.8	0.4	0.0	-0.9	0.7	0.0	-0.3	-3.2	-1.3	-4.9	-0.1	-0.1	-2.3	-5.6	-5.2

TABLE VII

IMPACT ON MISLEADING AND DETECTABILITY OF USING A RANDOM SELECTION OF CLASSIFIERS WITH VARYING NUMBER OF ITERATIONS IN CRAFTING ADVERSARIAL PERTURBATIONS (VALUES ARE THE VARIATION OF THE SCORE COMPARED TO THE SCORE OBTAINED WHEN USING AN ENSEMBLE OF CLASSIFIERS [6]). GREEN SHADING REPRESENTS AN INCREASE IN MISLEADING RATE OR A DECREASE IN DETECTABILITY. RED SHADING REPRESENTS A DECREASE IN MISLEADING RATE OR AN INCREASE IN DETECTABILITY. KEY – R18: RESNET18; R50: RESNET50; A: ALEXNET; DN: DENSENET161; T1: TOP-1 MISLEADING RATE; T5: TOP-5 MISLEADING RATE.

			Misleading ↑								Misleading with defense ↑								
# iterations Ensemble	# iterations Random	Classifier	R18		R50		A		DN		R18		R50		A		DN		Det. ↓
			T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	
20	20	R18+R50	-0.4	-1.9	-0.2	-0.9	-1.8	-2.1	-6.9	-8.0	-10.0	-14.7	-5.7	-8.6	-2.1	-1.6	-5.3	-7.9	9.3
		R18+A	-0.6	-3.8	-5.1	-5.7	-0.2	0.3	-4.7	-4.7	-12.3	-19.4	-4.1	-5.0	1.2	-2.5	-3.8	-4.8	13.6
		R50+A	-7.0	-8.2	-2.7	-9.6	-0.1	0.3	-5.8	-6.6	-6.8	-5.7	-13.2	-19.7	-1.1	-2.1	-6.5	-5.2	0.9
		R18+R50+A	-0.3	-1.8	-2.3	-8.2	-1.4	-3.7	-8.4	-10.2	-18.0	-27.2	-0.8	-0.7	-2.6	-6.2	-8.9	-11.1	4.9
20	40	R18+R50	0.0	0.7	0.0	0.7	36.6	65.4	8.7	11.1	0.6	1.8	7.4	11.0	1.3	5.4	9.7	11.6	-7.9
		R18+A	0.0	0.6	7.4	8.6	0.0	0.9	7.2	8.8	-3.3	-5.1	5.6	6.2	12.2	19.0	5.9	6.8	-3.4
		R50+A	6.7	10.0	0.1	1.4	0.0	0.5	8.7	10.1	5.8	8.3	-3.3	-5.0	8.9	18.8	5.1	7.2	-9.8
20	60	R18+R50+A	0.0	1.0	0.1	1.6	0.0	1.1	11.4	16.4	3.9	9.6	8.5	19.1	6.2	12.0	8.9	12.4	-28.7
60	60	R18+R50+A	0.0	0.2	0.0	0.0	-0.1	-0.4	6.6	10.7	15.2	28.1	21.2	37.0	10.4	18.7	13.1	18.3	-28.9

uses of a classifier, the random selection of classifier improves the misleading of unseen classifiers by 16.4 percentage points in the top-5 (when using three classifiers), and decreases the detectability by 28.7 percentage points. Moreover, when the ensemble uses the classifiers three times more than RP-FGSM, RP-FGSM still outperforms the use of ensemble by an average of 15.0 percentage points in terms of misleading rate for classifiers with defense and by 28.9 percentage points in detectability (Table VII, last row).

This ablation study confirms that randomly selecting classifiers and transformations can improve the misleading rate, with and without defenses, when attacks are evaluated on the same number of forward/backward passes of a classifier.

Figure 5 reports the effect of varying the parameter $\epsilon \in \{1, 2, 4, 8, 16, 32, 64\}$ (larger values produce lower quality images [12]). Image quality is reported as PSNR, and we attack single classifiers, i.e. ResNet18, ResNet50 and AlexNet, and also their combination. With seen classifiers, misleading rates above 95% occur with ϵ larger than 8 (with defenses) and 32 (without defenses). With unseen classifiers, the proposed attack performs similarly with and without defenses, thus supporting the use of transformations to make the adversarial images robust to defenses. The proposed strategy of randomly choosing a classifier at each iteration, when multiple classifiers are considered, increases the misleading rate with unseen classifiers as shown in the first two plots on the right column. The detectability rate does not show a clear relation with the parameter ϵ, as it remains between 40% and 80% for any ϵ when one classifier is attacked. When three classifiers are attacked the detection rate increases for $\epsilon \in \{1, 2, 4, 8\}$ and it establishes at around 80% for larger ϵ. Image quality monotonically decreases when ϵ increases. However, when multiple classifiers are used the PSNR does not significantly change.

F. Runtime analysis

We compare the runtime of each attack in an Ubuntu 18.04.3 server equipped with an NVIDIA Tesla V100 GPU, using a random subset of 300 images. The implementations are in Python using the PyTorch library [28]. Figure 6(a) shows that JSMA, CW and SparseFool are the slowest attacks taking on average 52.14, 9.87 and 4.42 seconds per image, respectively, when attacking ResNet50. DeepFool and FGSM-based attacks have a similar runtime performance with an average under 0.7 seconds per image. Figure 6(b) shows how the runtime of RP-FGSM increases when the number of attacked classifiers increases (similarly to DI-FGSM and E-FGSM). The major factor affecting the runtime is the forward/backward passes on the classifier.

VI. CONCLUSION

We presented RP-FGSM, a Robust and Private Fast Gradient Sign Method that is designed to mislead seen and unseen classifiers with and without known defenses. RP-FGSM has better performance than other state-of-the-art adversarial attacks for privacy protection in the Private Places365 dataset,

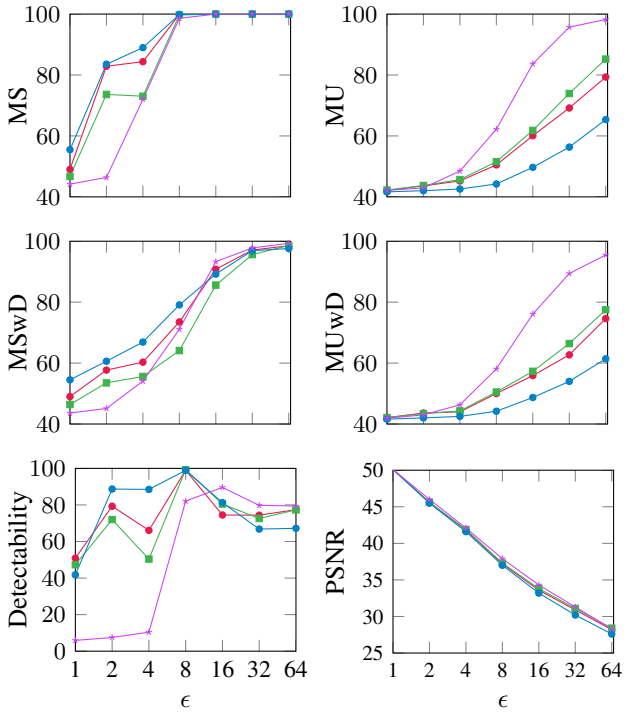


Fig. 5. Sensitivity analysis of targeted RP-FGSM when ϵ varies. The misleading rate is evaluated on top-1 on the test dataset. When using three classifiers to craft the adversarial images, they are evaluated with the most accurate classifier, ResNet50 (seen classifier) and DenseNet161 (unseen classifier). KEY – —: ResNet18; —: ResNet50, —: AlexNet; —: the three classifiers; MS: misleading a seen classifier; MU: misleading an unseen classifier; MSwD: MS with defense; MUwD: MU with defense.

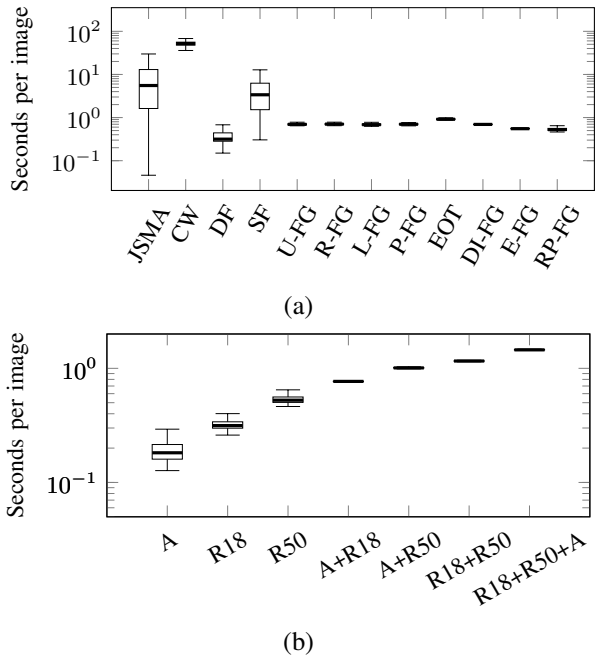


Fig. 6. Runtime analysis as average seconds per image. The test was performed on a random subset of 300 images. The horizontal line within the box shows the median; the bottom and top edges show the minimum and maximum values; and the lower and upper edges show the 25-percentile and 75-percentile, respectively. (a) Different adversarial attacks with ResNet50; and (b) proposed RP-FGSM when attacking a varying number of classifiers. KEY – JSMA: Jacobian-based Saliency Map Attack; CW: CarliniWagner; DF: DeepFool; SF: SparseFool; U-FG: Untargeted FGSM; R-FG: Random FGSM; L-FG: Least-Likely FGSM; P-FG: Private FGSM; EOT: Expectation Over Transformation; DI-FG: Diverse Input FGSM; E-FG: Ensemble FGSM; RP-FGSM: proposed attack; A: AlexNet; R18: ResNet18; R50: ResNet50.

especially for unseen classifiers and when defenses are applied. The key for this performance is the random selection of a defense transformation and a classifier at each iteration, which prevents the crafted perturbation from overfitting to a particular classifier or defense. As future work, we will extend the validation to tasks beyond scene classification.

ACKNOWLEDGMENT

We thank the Alan Turing Institute (EP/N510129/1), which is funded by the EPSRC, for its support through the project PRIMULA.

REFERENCES

[1] C. Y. Li, A. S. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 2502–2506.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, USA, May 2015.

[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbrücken, Germany, March 2016, pp. 372–387.

[4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the Symposium on Security and Privacy (S&P)*, San Jose, California, USA, May 2017, pp. 39–57.

[5] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July 2018, pp. 284–293.

[6] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, April 2018.

[7] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings of the Network and Distributed Systems Security symposium (NDSS)*, San Diego, California, USA, February 2018.

[8] Z. Liu, Q. Liu, T. Liu, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019, pp. 860–868.

[9] "Pixel privacy task, MediaEval 2018," http://www.multimediaeval.org/mediaeval2018, 2018, [Last accessed March 2020].

[10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, April 2018.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proceedings of the International Conference on Learning Representations (ICLR) Workshop Track*, Toulon, France, April 2017.

[13] C. Xie, Z. Zhang, J. Wang, Y. Zhou, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019, pp. 2730–2739.

[14] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, California, USA, June 2019, pp. 4312–4231.
[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, June 2016, pp. 2574–2582.
[16] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "SparseFool: a few pixels make a big difference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019, pp. 9087–9096.
[17] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Empirical study of the topology and geometry of deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, June 2018, pp. 3762–3770.
[18] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," *Journal of the Optical Society of America (JOSA)*, vol. 15, no. 8, pp. 2036–2045, August 1998.
[19] R. Shin and D. Song, "JPEG-resistant adversarial images," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS) workshop*, Long Beach, California, USA, January 2017.
[20] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 6, pp. 1452–1464, June 2018.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, June 2016.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, USA, December 2012, pp. 1097–1105.
[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR)*, Honolulu, Hawaii, July 2017, pp. 4700–4708.
[24] G. Dziugaite, Z. Ghahramani, and D. Roy, "A study of the effect of JPG compression on adversarial images," *arXiv preprint arXiv:1608.00853*, August 2016.
[25] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression," *arXiv preprint arXiv:1705.02900*, May 2017.
[26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[27] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1 – 21, January 2010.
[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proceedings of the Advances in Neural Information Processing Systems Workshop (NIPS-W)*, Long Beach, California, USA, December 2017.

Ricardo Sanchez-Matilla is a research assistant and a Ph.D. candidate with the Centre for Intelligent Sensing at Queen Mary University of London. He received his BSc and MSc degree in Telecommunication Engineering from the Universidad Autonoma de Madrid, Spain, in 2015. His research interests include computer vision, deep learning, and privacy.

Chau Yi Li is a Ph.D. candidate in the Centre for Intelligent Sensing at Queen Mary University of London. After receiving her BSc in Mathematics and BEng in Information Engineering from The Chinese University of Hong Kong in 2015 and 2016, she received her MSc in Computer Science from Queen Mary University of London in 2017. Her research interests include underwater image processing, deep learning and privacy.

Ali Shahin Shamsabadi is a PhD student in the Centre for Intelligent Sensing (CIS) in the school of Electronic Enginnering and Computer Science at Queen Mary University of London. His research interests are within the intersection of Machine Learning, Privacy and Image Processing. He aims to addresse the privacy risks in Machine Learning as a Service, which can be categorized in three research themes: privacy-preserving centralized learning, distributed learning and adversarial attacks for privacy protection.

Riccardo Mazzon received his BE in 2006 and MSc in 2009 in Computer Engineering from the University of Padova, Italy, and his PhD in Electronic Engineering from Queen Mary University of London (QMUL), UK, in 2013. Currently, he is a Research Manager at the Centre for Intelligent Sensing (CIS) and his research interests include video processing and analysis, camera networks and privacy.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He is Professor of Multimedia Signal Processing and the founding Director of the Centre for Intelligent Sensing at Queen Mary University of London (QMUL, UK), Turing Fellow at the Alan Turing Institute, the UK National Institute for Data Science and Artificial Intelligence, and Fellow of the International Association for Pattern Recognition. He is Editor-in-Chief of Signal Processing: Image Communication; Senior Area Editor for the IEEE Transactions on Image Processing; Chair of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee; and an IEEE Signal Processing Society Distinguished Lecturer.